Identifying Job Titles and the Boundaries between Job Ads

Enghin Atalay, Phai Phongthiengtham, Sebastian Sotelo, Daniel Tannenbaum[1]

Abstract

In this document, we explain how we identify the job title from an individual job ad and how we discern the boundaries between individual job ads.

Given a page of job ads, our next tasks are to discern the boundaries of individual vacancy postings and to identify each posting's job title. To perform these two tasks, we rely on regular patterns by which employers format their job ads. Many job ads begin with a job title, with the title appearing in upper case letters in a few-word line. Many job ads also contain a similar set of phrases at the end of each ad: "equal opportunity employer," "affirmative action employer," "send resume to," "in confidence to," and so on. In this document, we detail our algorithm to identify job ads' titles and their boundaries. We then discuss, using an example from our newspaper text, the performance of our algorithm.

The first step of our procedure identifies whether a particular line of text represents a job title. This is the case if one of two conditions are met. First, we search for lines for which all words are upper case and at least one of the words in the line of text correspond to a job title word.[2] Second, we search for lines in which the line is sufficiently short (there are at most three words on the line), one of the words of the text corresponds to a job title word, and the previous line of text did *not* represent a job title. If multiple lines of text are in all upper case, we combine these lines of text to form the job title. Having identified job titles, we demarcate this line of text as the boundary between job ads.

The second step of our procedure searches for other patterns which identify the end of a job ad. The comprehensive list of patterns which we search for are:

- "affirmative [any word] employer,"
- "in confidence,"

[1]Atalay and Phongthiengtham: Department of Economics, University of Wisconsin-Madison. Sotelo: Department of Economics, University of Michigan-Ann Arbor. Tannenbaum: Department of Economics, University of Nebraska-Lincoln. We acknowledge financial support from the Washington Center for Equitable Growth.

[2]The set of job title words is posted on the data library for this project. To produce this list, we retrieved the "Sample of Reported Job Titles" from the O*NET website. From this list, we construct a list of one-word personal nouns

- "submit resume,"
- "equal opportunity" or "equal opportunities,"
- "for further information contact,"
- a phone number, which is identified as a 7 digit or 10 digit number, with or without hyphens, or
- an address, which is identified as a number plus "ave," "blvd," "street," "hgwy," "road," or "drive."

It is likely that several of these phrases appear in succession at the end of job ads. To the extent that we recognize multiple such patterns in an individual job ad, our procedure will generate several short identified job ads. These short fragments of text will be discarded from our final data set. However, since these text fragments describe firms' contact information, and do not contain information related to the job's task content or skill requirements, discarding these fragments does not pose a problem for our analysis.

An Example

To make this procedure concrete, we present an example describing how our procedure identifies job ads' boundaries and their job titles. Figure 1 presents a snippet of Display Ad #226, from the January 14, 1979 Boston Globe.

In Figure 1, the first three lines of text are in all upper case. These lines of text are all combined to a single job title: "Medical Help Nuclear Radiologic Tech." Later on in the text, our procedure recognizes "Chest Physical Therapist," "Manager of Primary Care Programs," and "Medical Group Manager," among others, as job titles. However, even though "The Malden Hospital" appears in all upper case in a single short line, because this phrase does not contain a job-title-word, it is not identified as a job title.

Figure 2 contains the results from our exercise. In this figure, each ad is depicted as a four dimensional object. The first field gives the newspaper title, date, and page number. The second and third fields give the job ad number and the job title, while the final field gives the text within the job ad. As discussed above, the first two ads' job titles are recorded as "medical help nuclear radio logic tech" and "chest physical therapist." The phrase "an equal opportunity employer" appears in the second ad. Since this is a phrase we search for in demarcating the end of a job ad, our procedure identifies the end of the second ad here. The next line of text is (part of) an address: 41 Pa Hill Boston. As a result, it is also classified as the end of a job ad. Thus, job ad #3 is a short four-word ad, with no title. Since there is

Figure 1: Snippet of the January 14, 1979 Boston Globe, Display Ad #226

MEDICAL HELP \n NUCLEAR \n RADIOLOGIC TECH \n full time day po ition is available for registred or registry technician in our Nuclear Medicine department This position does require taking call \n CHEST \n PHYSICAL THERAPIST \n If you are or registry eligible \n Physical Trhrapist interested in Chest \n Therapy consider the New England Baptist Hospital Responsibilities will include providng chest therapy for Medical Surgical patients family teaching interdisciplinary inservice programs and more \n For more information please contact our Personnel department 738-5800 , Ext 255 . An Equal Opportunity Employer \n 41 Pa HII Boston \n MANAGER OF \n PRIMARY CARE PROGRAMS \n Children's Hospital Medical Center \n seeks dynamic creative individual to manage its Primary Care Programs including 24-hour Emergency Room Primary Care program the Massachusetts Poison information Center and \n Dental services This position requires 3-5 years experience with background in planning budgeting and managing \n health programs Masters degree preferred but additional experience may be substituted We offer salary commensurate \n with experience and fine fringe benefits package \n please forward resumes to Helena Wallace personnel office \n MEDICAL \n 300 Lonjwood Avenue \n MA 0211 \n REGISTERED \n REGISTRY ELIGIBLE OR \n immi ate available in our modern well- and fu ly accredited 173-bed general hospital Cheshire Hospilal is 80 miles from Boston and near skiing water sports hunting and fishing \n Apphcants must be registered registry eligible or NERT For further information please contact the Personrel department \n Cheshire Hospital \n 580 Court Street Keene NH 03431 \n equ ply MF \n MEDICAL GROUP MANAGER \n For North Shore group of 9 physicians Internal Medicine Radiclogy in physician-owned building Managerial and business skills required \n Knowledge of accounting essential Familiarity with care field desirable \n Apply to and include salary range expected Walter O'Donnell \n CAPE ANN MEDICAL CENTER \n Gloucester Mass 01930 \n 's 's \n immediate openings on rotating evening and night shifts for 's and 's for float pool assignments Experience in acuto nursing degree preferred for RN \n Apply PERSONNEL DEPARTMENT \n CAPE COD HOSPITAL \n Hyannis MA 02601 \n An qual \n MEDICAL RECORDS SUPERVISOR \n Challenging for capable professional to assume su responsib mly of Meaical Records department Involvement with all aspects medical records process with exception of ion Requires 2-3 years supervisory experience Candilate should be ART RRA registered or eligible for AMARAexam \n please send resume qnd requirements in confidence to Director of Personnel \n THE MALDEN HOSPITAL1 \n Hospital Road Malden Mass 02148 \n -1pwo rtu \n MEDICAL \n RECEPTIONIST \n Heavy manuscript typing gree inq patients and secretarial duties Kno of medical ro Yping 65 wpm \n Synd resume to Dobra Kiley -Davis 50 Binney Strett Boston MA 02115

Figure 2: Snippet of the January 14, 1979 Boston Globe, Display Ad #226

Globe_displayad_19790114_226|1|medical help nuclear radio logic tech|full time day po it ion is available for registered or registry technician in our nuclear medicine department this position does require taking call
Globe_displayad_19790114_226|2|chest physical therapist|if you are or registry eligible physical therapist interested in chest therapy consider the new England baptist hospital responsibilities will include providing chest therapy for medical surgical patients family teaching interdisciplinary in service programs and more for more information please contact our personnel department 738-5800 , ext 255 . An Equal Opportunity Employer
Globe_displayad_19790114_226|3||41 pa hii boston
Globe_displayad_19790114_226|4|manager of primary care program|children hospital medical center seeks dynamic creative individual to manage its primary care programs including 24-hour emergency room primary care program the Massachusetts poison information center and dental services this position requires 3-5 years experience with background in planning budgeting and managing health programs masters degree preferred but additional experience may be substituted we offer salary commensurate with experience and fine fringe benefits package please forward resumes to Helena Wallace personnel office
Globe_displayad_19790114_226|5|medical|300 lonjwood avenue
Globe_displayad_19790114_226|6||ma 0211 registered registry eligible or immi ate available in our modern well- and fu ly accredited 173-bed general hospital cheshire hospital is 80 miles from boston and near skiing water sports hunting and fishing applicants must be registered registry eligible or inert for further information please contact the person rel department cheshire hospital 580 court street keener nh 03431
Globe_displayad_19790114_226|7||equ ply
Globe_displayad_19790114_226|8|mf medical group manager|for north shore group of 9 physicians internal medicine radiology in physician-owned building managerial and business skills required knowledge of accounting essential familiarity with care field desirable apply to and include salary range expected Walter O'Donnell
Globe_displayad_19790114_226|9|cape ann medical center|Gloucester mass 01930
Globe_displayad_19790114_226|10||immediate openings on rotating evening and night shifts for and for float pool assignments experience in auto nursing degree preferred for
Globe_displayad_19790114_226|11||apply personnel department cape cod hospital
Globe_displayad_19790114_226|12||an equal
Globe_displayad_19790114_226|13|medical record supervisor|challenging for capable professional to assume su responsible ml y of medical records department involvement with all aspects medical records process with exception of ion requires 2-3 years supervisory experience can dilate should be art rra registered or eligible for amaraexam please send resume qnd requirements in confidence to Director of personnel
Globe_displayad_19790114_226|14|| the maiden hospital 1 hospital road alden mass 02148
Globe_displayad_19790114_226|15||-1pwo rt u
Globe_displayad_19790114_226|16|medical receptionist|heavy manuscript typing agree inq patients and secretarial duties Kano of medical ro y ping 65 wpm synod resume to do bra kiley -Davis 50 Kinney street Boston ma 02115

Notes: The first field within each line gives the source (Botson Gobe Display ads), day (January 14, 1979), and page number (226) of the advertisements. The second field gives the ad number within this page of ads. The third field gives the job title we have extracted, if any. The final fields give the text of the ad.

no job title, and also because this ad is exceedingly short — we exclude any ad, even those with job titles, if it has fewer than 15 words — this ad will be dropped from our final data set.

Combining Similar Job Titles, Flagging Spurious Job Titles

In a final step, after finding the boundaries between individual ads and identifying the set of words which plausibly refer to a job title, we i) combine job titles which are extremely similar to one another, and ii) flag what could be spurious job titles.

Regarding (i), we have combined near-identical job titles: removing words relating to salaries, locations, or hours that previously appeared in the job title field; standardizing acronyms; combining male-specific and female-specific job titles (e.g., host versus hostess).

Regarding (ii), our algorithm thus far identifies certain groups of words as referring to job titles, but more likely should not be classified as such. These include "vets" or "parker." "Vets" could conceivably indicate a veterinarian position, but which in the job ad text were solicitations for applications from veterans of recent wars. "Parker" could refer someone who parks cars, but is actually the name of an employment agency. We construct a new variable description_new_miss, which we set equal to 1 if the job title field is likely to be spurious.